PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
G06F 17/30, 17/50, G01N 33/00

A1

(11) International Publication Number: **WO 97/14106**

(43) International Publication Date: 17 April 1997 (17.04.97)

(21) International Application Number: PCT/US96/16196

(22) International Filing Date: 10 October 1996 (10.10.96)

(30) Priority Data:

08/542,642	13 October 1995 (13.10.95)	US
08/550,847	31 October 1995 (31.10.95)	US

(71) Applicant: TERRAPIN TECHNOLOGIES, INC. [US/US]; 750-H Gateway Boulevard, South San Francisco, CA 94080 (US).

(72) Inventors: VILLAR, Hugo, O.; 7706 Hazelnut Drive, Newark, CA 94560 (US). BONE, Richard, G., A.; 1553-10th Avenue, San Francisco, CA 94122 (US).

(74) Agents: MURASHIGE, Kate, H. et al.; Morrison & Foerster, L.L.P., 2000 Pennsylvania Avenue, N.W., Washington, DC 20006-1888 (US).

(81) Designated States: AU, CA, JP, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published

With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: IDENTIFICATION OF COMMON CHEMICAL ACTIVITY THROUGH COMPARISON OF SUBSTRUCTURAL FRAGMENTS

(57) Abstract

A process for preparing a database of molecule fragments contained by a molecule, the process including the steps of (1) selecting a molecule comprising molecule fragments; (2) identifying substantially all molecule fragments within the molecule; (3) counting the occurrence of each unique molecule fragment identified in step 2; and (4) storing in a computer-readable storage medium information correlating molecule fragment count with molecule fragment identity. The invention also relates to a computer-implemented process for identifying a molecule likely to have a desired trait, the process including the steps of (1) selecting a test molecule comprising molecule fragments; (2) identifying substantially all test molecule fragments within the molecule; (3) counting the occurrence of each unique test molecule fragment identified in step 2; (4) comparing test molecule fragment counts from step 3 with fragment counts for identical fragments from a plurality of molecules having known activity; and (5) presenting the results of the comparing step as an output. The invention also includes a data processing system implementing these processes and computer-readable media having recorded thereon databases containing standardized representations of substantially all fragments of one or more molecules.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AM	Armenia
AT	Austria
AU	Australia
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CS	Czechoslovakia
CZ	Czech Republic
DE	Germany
DK	Denmark
EE	Estonia
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GE	Georgia
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IT	Italy
JP	Japan
KE	Kenya
KG	Kyrgystan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakhstan
LI	Liechtenstein
LK	Sri Lanka
LR	Liberia
LT	Lithuania
LU	Luxembourg
LV	Latvia
MC	Monaco
MD	Republic of Moldova
MG	Madagascar
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
MX	Mexico
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SG	Singapore
SI	Slovenia
SK	Slovakia
SN	Senegal
SZ	Swaziland
TD	Chad
TG	Togo
TJ	Tajikistan
TT	Trinidad and Tobago
UA	Ukraine
UG	Uganda
US	United States of America
UZ	Uzbekistan
VN	Viet Nam

IDENTIFICATION OF COMMON CHEMICAL ACTIVITY THROUGH COMPARISON OF SUBSTRUCTURAL FRAGMENTS

Technical Field

The invention relates to the study of structure-activity relationships in small molecule ligands. In particular, it concerns methods and devices to identify and correlate structural features of molecules with other molecules and/or libraries of molecules in order to identify the likely activity of the molecules in question. The invention employs systematic identification of structural fragments and a counting means to provide correlations.

Background Art

Interaction between ligands and ligates, accounting for biological activity of the ligand or ligate, depends on the presence in the ligand of certain functional groups and their arrangement. Many attempts have been made to associate particular structural features with functionality, such as pharmacological activity, receptor binding ability, catalytic function, and the like. One approach has focused on the presence of certain common structural features, identifiable by visual inspection of structural formulas, which seem to be shared by molecules with similar pharmacological activities. For instance, central nervous system active drugs often contain piperidine rings. Benzodiazepine rings also are present in a variety of pharmaceuticals. Computer modeling of three-dimensional structures to ascertain the features of active sites has also been used as a basis for the design of small molecule mimics for ligands or ligates. Reactivity fingerprints with panels of reagents, such as antibodies or paralogs having various specificities have also been used to identify molecules which have specific biological binding specificities. Structural similarities among successful molecules identified by these methods have also been correlated by inspection of the relevant formulas.

Nevertheless, the key functional groups of interest and their arrangement are not necessarily evident from inspection if there is no systematic means for their identification and subsequent correlation with biological or other activity. This problem is recognized in the literature. For example, Martin, E. J., et al., in J Med Chem (1995) 38: 1431-1436 disclose a specific method for characterizing a particular class of molecules so that such correlations can be made. These authors identified 15-20 specific properties for each monomer composing the side chains of a series of oligo (N-substituted) glycine "peptoids." These descriptors such as lipophilicity, various topological indices, and functionality descriptors can be represented in the form of "flower plots" which show (on a circularized Y-axis) the extent of each property which has been assigned a value on the X-axis. Similarity of the side chains of the peptoids can then be readily determined by comparing their flower plots. Numerous different molecules can share the same flower plot.

Applying similar concepts to broader classes of compounds requires definition of functional groups by some means other than a "side chain" of a polymer backbone. Sello, G., "A New Definition of Functional Groups and a General Procedure for Their Identification in Organic Structures," 114 J.Am.Chem.Soc. 3306-3311 (1992), states that the recognition of functional groups in organic molecules is important both for their handling (*e.g.*, in organic reactivity modeling) and for their storage and retrieval in reaction databases. Sello describes two fundamental approaches: (1) the choice of a set of fundamental functional groups that permits the recognition of the functional groups in a given molecule via an accurate comparison between its atomic groups and the functional group set, and (2) the definition of a set of rules, listing the necessary requisites of a functional group, that can be applied to a given molecule furnishing its functional groups. Sello defines a "functional group" as "a set of sufficiently important connected atoms, in which the importance is always decreasing from the central atom towards the peripheral atoms", where "importance" is equivalent to "contribution to the electronic stabilization energy of the molecule." Sello, however, does not discuss the identification or comparison of any sub-molecular group other than the functional groups he defines. More importantly, he does not provide a means for providing a complete enumeration of sub-molecular groups to be evaluated as candidates for inclusion in a list of functional groups as he has defined them.

The problem of enumeration of unique molecular structures has also been recognized in the literature. Liu, X., et al., "The Graph Isomorphism Problem," 10 J. Comp. Chem. 1243-51 (1991), describe the use of topological indices as a threshold test for the comparison of two molecules, followed by the use of isomorphism testing if the topological indices match, as a method of determining whether one molecule is unique from another. Liu et al. do not, however, apply the technique to molecule fragments, nor do they use the result of their comparisons to determine the likely attributes of one of the molecules or to derive libraries or other databases useful for such comparisons.

The disclosures of each of these prior art publications is incorporated herein by reference.

The present invention provides a complete systematic classification of functionality or activity based on particular topological characteristics of fragments of small molecules, with utility in a wide variety of structure/function comparisons. These data further provide a basis for construction of efficient combinatorial libraries covering essentially all of chemical space or focused on variations on particular themes to fine-tune selection of molecules with desired activity.

Disclosure of the Invention

This invention relates to a process for preparing a database of molecular fragments contained by a molecule, the process including the steps of (1) selecting a molecule; (2) identifying substantially all sequentially attached molecular fragments within the molecule; (3) counting the occurrence of each unique molecule fragment identified in step 2; and (4) storing in a computer-readable storage medium information correlating molecule fragment count with molecule fragment identity.

The invention also relates to a computer-implemented process for identifying a molecule likely to have a desired trait, the process including the steps of (1) selecting a test molecule comprising molecule fragments; (2) identifying substantially all molecule fragments within the test molecule; (3) counting the occurrence of each unique test molecule fragment identified in step 2; (4) comparing test molecule fragment counts from step 3 with fragment counts for identical fragments from a plurality of molecules having known activity; and (5) presenting the results of the comparing step as an output.

The invention also includes a data processing system implementing these processes and computer-readable media having recorded thereon databases containing standardized representations of substantially all fragments of one or more molecules.

The invention is described in more detail below with reference to the drawings.

Brief Description of the Drawings

Figure 1 is a flow chart showing a process for assembling a library of fragmented molecules.

Figure 2 is a flow chart showing a process for comparing a test molecule with a library of fragmented molecules of known activity.

Figure 3 is a table showing all distinct subgraphs, listed by vertex label, for the skeleton methyl-cyclopentane.

Figure 4 is a flow chart showing a method of identifying and enumerating fragments of molecules.

Figure 5 is a flow chart showing a preferred method of identifying and enumerating molecule fragments, including a method of identifying duplicate fragments.

Figure 6 shows the distinct moieties for 3-methyl-tetrahydrofuran, along with their respective frequencies.

Figure 7 shows a single moiety from Figure 6 along with its positions of occurrence in the original molecular skeleton.

Figure 8 is a three-dimensional bar graph showing the number of distinct subgraphs, *S(N,m)*, of size *m*, as a function of chain length *N* for straight-chain molecules.

Figure 9 is a three-dimensional bar graph showing the number of distinct subgraphs, *S(N,m)*, of size *m*, as a function of ring size *N* for simple single-ring molecules.

Figure 10 shows nine isomers of heptane.

Figure 11 is a table showing the subgraph counts, *S(7,m)*, and moiety counts, *M(7,m)*, for each heptane isomer *N=7* shown in Figure 10.

Figure 12 is a table showing moiety counts, *M(N,m)*, for branched heptamine isomers where the numeric label denotes the position of the N-atom substituent in each respective skeleton labelled b-i in Figure 10.

Figures 13(a) and 13(b) are tables showing numbers of distinct subgraphs, *S(N,m)*, and moieties, *M(N,m)*, for the 20 naturally-occurring amino acids, ordered by size.

Figure 14 is a histogram showing distribution of moieties of all 20 amino acids from Figure 13(b) compared to the distribution of fragments from Tryptophane alone.

Figure 15 shows a molecule library made up of multiple molecules.

Figure 16 shows a test molecule.

Figure 17 is a table showing the results of the use of the invention to compare fragments from a test molecule with those in a library of similar molecules.

Figure 18 shows the fragments identified in the fragment comparison process as present in more than half of the library molecules and also either present or absent in the test molecule.

Methods of Carrying Out the Invention

In its most general terms, the invention provides a way of identifying a molecule likely to demonstrate certain desirable (or undesirable) activities, or put another way, the invention identifies the likely activity of a given test molecule. The desirable activities could be any measurable or calculated property of the compound, either biological, chemical or physical. The likelihood that any such activity is possible is inferred from a comparison of substructural fragments of the molecule with fragments of other molecules whose activities are known. A "fragment" is a covalently connected set of atoms, whether or not the arrangement could actually stably exist independently of the remainder of the molecule. Once its likely activity has been identified, the fragment can be assembled computationally with other fragments into discrete chemical structures, corresponding compounds can be synthesized by combinatorial chemical operations or other synthetic procedures.

The preferred method of this invention is shown in flowchart form in Figures 1 and 2. Figure 1 is a process for assembling a collection of fragments from a library of molecules of known activity for any particular function. The library can be chosen to contain molecules having one or more desired activities. One important library is made up of molecules chosen at random from the set of all known available chemicals without consideration for the particular activity of each molecule.

Figure 2 shows a process for comparing a fragmented test molecule of unknown activity with a library of fragmented molecules of known activity to identify the activity the test molecule is likely to have. The library of molecule fragments and their counts used in the process of Figure 2 can be assembled using the process of Figure 1 or another process.

Step 10 of Figure 1 selects a molecule having known calculated or measured activity. In step 12, the fragments of the molecule are identified in a unique manner. In step 14, the number of times each fragment occurs within the molecule is counted. In step 16, the fragment count is stored in a manner associating fragment count with fragment identity. Steps 10-16 may be repeated as often as desired to assemble a collection of fragments from a library of molecules having known activity.

Steps 20-24 of Figure 2 repeat the process of Figure 1 for a test molecule of unknown activity. Thus, the molecule is fragmented in step 22 and the fragments are counted in step 24. In step 28, the fragment counts are compared to the counts for identical fragments occurring in one or more molecules of known activity to identify the likely activity of the test molecule.

As shown in block 26 of Figure 2, the results of the fragment counting of step 24 may be stored in a manner associating fragment count with fragment identity before comparing the test molecule fragment counts with counts for known molecules in step 28. The process of this invention may also start with step 26, *i.e.*, with a pre-stored fragment list that is compared with fragment counts from a stored library of fragmented molecules of known activity, as in step 28.

The library or libraries of fragmented molecules of known activity to which the fragmented test molecule is compared may be chosen for the specific purpose of such comparisons. These libraries can include (i) a library of fragmented molecules having toxic properties; (ii) a library of fragmented drug molecules having one or a set of desired pharmacological activities (such as central nervous system active drugs, including opiates, benzodiazepine receptor ligands, neuropeptide receptor ligands, neurotransmitter receptors or their transporters); (iii) a library of fragmented random molecules; or (iv) a library of fragmented molecules specifically selected for common activity (or a range of activity values, such as low activity to high activity), a specialized example of which is common affinity fingerprints, as described in U.S. Patent No. 5,300,425 and pending U.S. patent applications S.N. 08/308,813 (filed 9/19/94); S.N. 08/177,673 (filed 1/6/94); and S.N. 08/477,132 (filed 6/7/95), the disclosures of which are incorporated herein by reference. Comparison of the fragmented test molecule with multiple libraries (or with a single multi-purpose library containing molecules of diverse, and known, activity) helps identify the likely activity of the molecule under analysis.

The preferred manner of implementing this invention is through use of a software program running on a digital computer yielding a physical database of actual utility. The program may be run on any data processing system, such as a computer, having sufficient computational power and memory. For example, we have run a version of the program successfully on an Indigo workstation (Silicon Graphics, Inc.).

The molecule to be fragmented for either the process of Figure 1 or the process of Figure 2 must be presented in a consistent format that can be recognized and manipulated by a computer. Many different formats have been proposed. The preferred embodiment of our invention uses the two-dimensional CTfile format developed by Molecular Design Limited (MDL), as described in

Dalby, A., *et al.*, "Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited," 32 J. Chem. Inf. Comput. Sci. 244-55 (1992), the disclosure of which is incorporated herein by reference. Thus, the preferred input to step 10 of Figure 1 or step 20 of Figure 2 is a MOLfile of a molecule as described in the Dalby article. Other two-dimensional molecule file formats may be used, however, without departing from the scope of the invention. *See, e.g.*, D. Weininger, "SMILES, A Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules," 28 J. Chem. Inf. Comput. Sci. 31-36 (1988). In addition, the two-dimensional molecule file format used in the practice of this invention may be derived from a suitable three-dimensional molecule file format.

In general, the description of a molecule in a two-dimensional (2D) chemical database is sufficient to carry out an enumeration of all its substructures. The standard representation of a molecule at the 2D-level is as a list of atom types and a connection table, *i.e.*, a list of pairs of atoms which are bonded to one another with a specified bond type. Other representations are possible, of course.

As is normally done for 2D-databases, hydrogens are omitted. Restricting the representation to "heavy atoms" makes the problem more manageable in most cases and does not result in loss of information because the knowledge of bond-types between the heavy atoms and standard assumptions of their respective valencies allows regeneration of full structural formulae for both parent molecule and its fragments.

From the connection table it is computationally straightforward to construct the molecular graph, and techniques of graph theory may then be applied. The *N* vertices (nodes) of the graph represent the atoms and are identified both by numerical labels (1, 2,..., *N*), and their atom types (C, O, N, S, *etc.*). Note the technical distinction between the atom "label" (which derives from an arbitrarily chosen numbering scheme) and the atom "type" which defines the chemistry of the problem. The "size" of a molecule may be quantified by the number *N* itself. The *E* edges of the graph represent the bonds between pairs of atoms and carry integer "weights" corresponding with chemists' conventional understanding of bond-order (single, double or triple in most cases). Hydrogen bonds are not considered. The "connectivity", *d*, of a given atom is the number of other atoms to which it is bonded, irrespective of bond-type. The graph is said to be "connected" if it is possible to trace paths along consecutive edges from every vertex to every other. All the graphs and their subgraphs dealt with here are assumed to be connected corresponding to discrete covalently bound molecules and molecule fragments, respectively.

The first step in the analysis is to generate all connected subgraphs of the "parent" molecular graph which contain distinct sets of vertices by label. A connected subgraph corresponds to a molecular fragment (which need not be capable of independent existence in its own right); at this stage it is uniquely defined by the list of atom labels which comprise it. Therefore, two subgraphs of the same size are distinct if they differ by at least one vertex label. For a given molecule with *N* atoms, the number of distinct subgraphs of size *m* will be designated by *S(N,m)*, where *m* ranges from 1 (single atoms) to *N* (the parent molecule).

By way of example, for the skeleton of methyl-cyclopentane, all the distinct subgraphs are listed by vertex label in Figure 3. The labelling of the atoms of the skeleton is arbitrary and purposefully does not follow any usual convention because the results of significance are independent of the choice of labels. At the bottom of the figure are the totals for each size. Note that, at this stage, the number of distinct subgraphs, *S(N,m)*, of each size, *m*, which are found is independent of the atom-types and bond-types and, of course, is also independent of the way in which the atoms are labelled. In other words, the numbers and types of subgraphs of each size is dictated solely by the individual connectivities of the atoms, *i.e.*, the molecular topology. In Figure 3, then, 2-methyl-tetrahydrofuran and 3-hydroxy tetrahydrofuran would give identical distributions of subgraphs. Such groups of molecules will be called "iso-skeletal." The utility of *S(N,m)* is that it offers a mathematical representation of atom-connectivities in a molecular structure which is irrespective of the atom types which are present but which is distinctive of certain topologies.

It should be noted that any pair of atoms which are bonded to one another in the parent molecule retain that bond in any fragment which contains them both. Therefore, there is one 5-atom subgraph which contains all 5 ring atoms; *i.e.*, "chains" of all 5 ring atoms are not found. Such a result is a consequence of defining subgraphs by sets of vertex-labels rather than sets of edge-labels which would be an equally valid, but distinct, approach. The rationale behind this choice is that a "substructure" is envisaged to be a set of atoms and the bonds between them extracted from the parent molecule, *in toto*, without considering the often many possible alternative arrangements of bonds which leave those atoms connected.

In order to incorporate the chemical attributes of the parent molecule into the purely topological information contained in the subgraphs, the iso-skeletal but chemically distinct subgraphs must be distinguished from one another. There are many levels at which this may be carried out. In the implementation described here, the subgraphs are distinguished only by the atomic numbers of the atoms involved and the bond types of the bonds between them. Therefore, at this level, all carbon atoms are equivalent to one another, as are all oxygens, *etc*. It is only the types and numbers of bonds that they form which differentiates them. For example, a carbonyl carbon can only be distinguished from an aliphatic carbon in moieties which contain the carbonyl oxygen atom also. Alternatively, atoms may be distinguished by their local environments, in which case, for example, carbon atoms have a number of types, such as aliphatic, aromatic, carbonyl, *etc*. It is also possible to use some other property, *e.g.*, charge or hydrophobicity, to distinguish atoms. A substructure defined in any of these ways will be referred to as a "moiety", in order to avoid confusion with other terms such as functional group or substructure. The term "fragment" is more general and can be used to mean either subgraph or moiety.

The distinct moieties are stored in different files according to their size and it is, of course, only necessary to compare moieties of the same size. Figure 4 is a flow chart showing a preferred way of identifying and enumerating moieties using a computerized data processing system. As in the embodiments discussed above, the process of this preferred embodiment begins at step 30 with the molecules stored in a computer-readable storage medium within the data processing system in a standardized form (such as the MDL MOLfile format) with substantially all chemical ambiguities resolved. Since it can be readily accomplished by a variety of means known to practitioners in the art of computational chemistry, the standardization process is not part of this invention.

In step 32, the molecule fragments are arranged according to fragment size, *i.e.*, in a two-atom fragment list, a three-atom fragment list, and so on, depending on the processing and memory limits of the computer system being used. In the preferred embodiment, the list starts with two-atom fragments and proceeds to three-atom fragments only after the two-atom list is exhausted and continues with fragments of increasing size. The list stops when the fragment size is the same as the molecular size or the limits of the computer system have been reached. While proceeding in order of increasing fragment size increases the efficiency of the comparison process, this step is optional.

In steps 34 and 36, a fragment is selected from the list and the system determines whether the fragment differs from any fragment previously stored in a fragment table within a computer-readable storage medium within the data processing system. This data storage medium can either be associated with, or independent from, the medium in which the standardized molecule information is stored.

The initial fragment selected at the start of the process will be unique, of course, since no fragments will have yet been stored in the fragment table. The process therefore proceeds to step 38, where the fragment identity is stored in the fragment table and the counter for that fragment is set at 1.

If there are more fragments in the molecule's fragment lists, the process returns to step 34 to select the next fragment. If this next fragment is the same as a fragment previously stored in the fragment table, the counter for that fragment is incremented at step 42, and the process returns to step 34 to select another fragment if there are any more fragments in the fragment lists. If the selected fragment differs from the fragments already stored in the table, the selected fragment is added to the table and its counter is set at 1.

The result of this process is a computer-readable database of one or more molecules listing the molecule fragments identified in the molecule and the number of times each fragment occurs in the molecule. This database can be displayed on a computer display, printed on a printer, or stored in a digital memory such as a ROM, RAM, magnetic tape or optical disk. Multiple known fragmented molecules may be compiled into a library to be used for comparison with a fragment list from a test molecule, as described above with reference to Figure 2. Alternatively, the fragmentation procedure may be performed on known and test molecules together without any intermediate storage or display of fragment count information.

The comparison of each fragment with those already stored can be a computationally intensive process. The preferred embodiment of our invention therefore uses a two-step approach to the fragment comparison process. The first step calculates an easy-to-compute index and stores it with each fragment. This index, however, is not necessarily unique; it is possible for more than one fragment to share the same index. The second step of the process, therefore, is to determine whether two fragments sharing the same index are identical through the use of a graph isomorphism algorithm. By using the computationally-intensive graph isomorphism comparison only in cases where a fragment's index matches another fragment's index, this approach minimizes computation time during the fragment identification and counting process.

An example of the use of this two-step process is shown in Figure 5. In steps 52 and 54 of Figure 5, topological indices are calculated for all fragments in the list of standardized fragments, and the fragments are arranged according to size. The preferred definition for topological indices is described in Hall, L.H., and Kier, L.B., "Determination of Topological Equivalence in Molecular Graphs from the Topological State," 9 Quant. Struct.-Act. Relat. 115-131 (1990), the disclosure of which is incorporated by reference. Other topological indices may be used without departing from the scope of our invention. *See, e.g.*, Randic, M., "On Characterization of Molecular Branching," 97 J. Am. Chem. Soc. 6609-15 (1975); Hosoya, H., "Topological Index: a Newly Proposed Quantity Characterizing the Topological Nature of Structured Isomers of Saturated Hydrocarbons," 44 Bull. Chem. Soc. Japan 2332-39 (1971); Schultz, H.P., "Topological Organic Chemistry. 1. Graph Theory and Topological Indices of Alkanes," 29 J. Chem. Inf. Comput. Sci. 227-28 (1989); Wiener, H., "Structural Determination of Paraffin Boiling Points," 69 J. Am. Chem. Soc. 17-20 (1947); Balaban, A.T., "Highly Discriminating Distance-Based Topological Index," 89 Chem. Phys. Lett. 399-404; Platt, J.R., "Influence of Neighbor Bonds on Additive Bond Properties in Paraffins," 15 J. Chem. Phys., 419 (1947); Bertz, S.N., "The First General Index of Molecular Complexity," 103 J. Am. Chem. Soc. 3599-3601 (1981); Bonchev, D., et al., "Information Theory, Distance Matrix and Molecular Branching," 67 J. Chem. Phys. 4517-33 (1977).

In steps 56 and 58, a fragment is selected from the list and its topological index is compared to the topological indices of fragments already stored in the fragment table. If the topological index of the selected fragment is unique, the fragment is stored in the fragment table and its counter is set at 1.

If, however, the selected fragment's topological index matches the topological index of a fragment already stored in the fragment list, the two fragments (*i.e.*, the selected fragment and the stored fragment having the same index as the selected fragment) are compared with a graph isomorphism algorithm. If the two fragments are isomorphic (*i.e.*, are identical), then the fragment's counter is incremented by 1, as in step 64.

On the other hand, if the selected fragment does not match any of the fragments in the table by the graph isomorphism method, the fragment is a new fragment and must be stored as such in the fragment table in step 60. This process of using a threshold comparison of a first, and less computationally intensive, representation of the fragments (*e.g.*, the topological indices) before using a second, more accurate but more computationally intensive, representation of the fragment reduces the computational load dramatically.

The preferred graph isomorphism method is a version of Ullmann, J.R., "An Algorithm for Subgraph Isomorphism," 23 J. Assoc. Comput. Mach. 31-42 (1976), the disclosure of which is incorporated herein by reference. Other approaches may be used, of course. *See, e.g.*, Barnard, J.M., "Substructure Searching Methods: Old and New," 33 J. Chem. Inf. Comput. Sci. 532-38

(1993); Corneil, D.G., et al., "An Efficient Algorithm for Graph Isomorphism," 17 J. Assoc. Comput. Mach. 51-64 (1970).

Since most molecules of interest have only a few non-carbon atoms, it is usually true that the number of distinct moieties is much less than the number of distinct subgraphs for any molecule. The numbers of distinct moieties of size *m* for a molecule of *N* atoms is designated by $M(N,m)$, where again *m* takes value between 1 (so that $M(N,1)$ represents the number of distinct atom-types in the molecule) and *N*.

The distinct moieties for 3-methyl tetrahydrofuran are shown in Figure 6 along with their respective frequencies. The numbers of distinct moieties for each respective size are shown at the bottom of the figure. In Figure 7, a single moiety is shown along with its positions of occurrence in the original molecular skeleton. As molecules and fragments get larger, reliable recognition of moiety (as either presence, absence or a simple count) becomes increasingly difficult to perform manually.

The following examples illustrate these principles for single molecules. As shown below, fragment counts have distinctive patterns for different classes of molecule.

Example 1.

For straight-chain molecules, the subgraphs are also all simple chains, ranging in size from single vertices up to the size, *N*, of the molecule itself. Their numbers are given by the simple linear relationship:

$$(1) \qquad S(N,m) = N - m + 1.$$

But, because all the atom-types are the same, there is no chemical distinction between any subgraph and any other of the same size, so there is only one distinct moiety for each size. Relationship (1) is shown for straight chains up to length 12, in Figure 8.

Example 2.

For simple, unsubstituted, single-ring systems in which all bonds are equivalent, (*e.g.*, cyclopropane, benzene, etc.) the numbers of subgraphs and moieties also follow very simple relations. For a ring of *N* atoms, there are exactly *N* subgraphs for each size *m* less than the ring-size itself, a result which is displayed in Figure 9 for monocyclic systems with up to 12 atoms. All such subgraphs are simple chains of length *m*, and for given *m*, they are identical with one another so that, just as with simple chains, the number of distinct moieties is again exactly 1 for each size.

Comparison of Figures 8 and 9 is instructive. The form of *S(N,m)* is qualitatively different for rings from those for the chain molecules. Simply by inspection it is possible to see that the functions *S(N,m)* for families of molecules with distinct topologies have very different characteristics. Since, for example, whether a molecule is ring-like or chain-like is likely to be a relevant property for understanding its likely activity, *S(N,m)* emerges as a powerful tool for quantifying intuition.

Example 3.

Figure 11 presents both subgraph counts, *S(N,m)*, and moiety counts, *M(N,m)*, for the 9 isomers of heptane listed in Figure 10. In order to investigate the effect of single hetero-atom substitution on branched acyclic systems, each symmetry-unique carbon of the nine isomers of heptane is replaced with a nitrogen atom, which leads to examples of primary, secondary, tertiary and quaternary amines. The structures are listed in Figure 12 (labeled **a** to **i** to correspond to Figure 10). They are grouped according to their skeletal form (in descending order of longest chain) in the order of the heptane isomers in Figure 11; their numbering in Figure 12 corresponds with the number of the nitrogen atom which is substituted for the carbon atom in each case, Figure 10.

While the numbers of distinct moieties for the heptamines is greater than that for the corresponding iso-skeletal heptanes, the overall distributions, *M(N,m)*, are still rather flat. For these molecules, the moiety diversity introduced by single heteroatom substitution is comparable to that introduced by branching, relative to an unbranched, unsubstituted molecule with the same number of atoms. This result implies that an increase in moiety diversity may be achieved both by introduction of branched centers and by insertion of atom types which differ from those which make up the main skeleton. The quantity *M(N,m)* serves to characterize the underlying chemical complexity of a molecule: large moiety counts may arise from complex topologies as well as the presence of distinct atoms. Molecules with high moiety counts have structures which permit a wide range of interaction types with external influences, for example, other molecules or biological receptors.

Example 4.

The fragment distributions in the 20 naturally occurring amino acids provide a good example of the insights to be gained from moiety enumeration. The molecules range in size from 5 heavy atoms (glycine) to 15 (tryptophane). Figures 13(a) and 13(b) give the distributions, *S(N,m)* and *M(N,m)*, respectively, for the 20 molecules, in ascending order of the number of heavy atoms.

Several of the amino acids are iso-skeletal and so have identical entries for *S(N,m)*. They are: cysteine and serine; threonine and valine; asparagine, aspartic acid and leucine; glutamic acid and glutamine.

The distributions of *S(N,m)* differentiate the amino acids into two types according to whether they contain a ring in their structure or not. The acyclic molecules (alanine, cysteine, threonine, asparagine, isoleucine, methionine, glutamic acid, lysine and arginine -- and their respective isoskeletal partners) have more atoms $m=1$ than bonds $m=2$ so that *S(N,m)* initially dips on increasing *m* from 1. But there is a subsequent "peak" in the distribution at larger *m*. This form for *S(N,m)* is characteristic of branched acyclic skeletons. By contrast, the amino acids which contain a ring (i.e., proline, histidine, phenylalanine, tyrosine and tryptophane) have a simple rise and subsequent decline in *S(N,m)* as *m* increases.

In Figure 13(b), the influence of the presence of heteroatoms on the moiety counts is observed clearly. The presence of more than one different type of hetero-atom means that there are many more distinct moieties than are possible respectively in isoskeletal hydrocarbons or singly-substituted analogs. It can also be seen that the larger the fragment size, the closer is the number of distinct moieties to the number of subgraphs themselves. In the limiting case in which every atom in a skeleton is distinct, *S(N,m)* and *M(N,m)* would be equal for all *m*.

Also conspicuous is the small total number of distinct moieties, *i.e.*, chemical substructures, which are present among the set of amino acids. The moieties number a few hundred at most, to be compared with over a thousand which could be obtained for, *e.g.*, cages or fused polyhexes of comparable numbers of atoms. A "binning" exercise was carried out in which the 20 amino acids were pooled and the total numbers of distinct moieties of each size from the whole set was counted. They are histogrammed in Figure 14. Note that the overall distribution is effectively dominated by the numbers of moieties which can be obtained from the largest molecule, tryptophane. The significantly larger number of moieties that tryptophane contains may account for its over-representation at the binding sites of anti-bodies and other proteins. It can be argued that the large number of substructures present in it provide a larger assortment of possible interactions with ligands than do the smaller amino acids.

There are important consequences of these results for the design of the scaffolds which underpin combinatorial chemistry methods. A desirable property of a chemical library is that it should contain high diversity. In a very crude way, this may be simply measured by a count of the numbers of distinct moieties which are present. Such a count measures the basic composition of the molecular constituents at the 2-D level, from which it can be extrapolated that at the 3-D level, the likelihood of matching a given pharmacophore also depends on the number of distinct substructures which are present. From the analysis of fragments provided by the techniques presented here, it is apparent that the higher the average connectivity of the atoms in the basic skeleton, the greater the possibility that the range of possible fragments is widely sampled. It is thus now possible to design molecules with useful or unusual fragment distributions.

Because real fragments of size 8-12 atoms are able to span the typical distances between receptor-interaction points, it is likely to be important to maximize the occurrence of substructures in this size range; smaller fragments are unable to achieve significant geometries and larger ones become unwieldy. Substructure counting has provided objective evidence that branched acyclic molecules and cage-like and fused-ring scaffolds are useful because they compactly give rise to maximal substructural representation. Branched acyclic molecules have the disadvantage of high conformational flexibility. One example of this general principle has been provided by Rebek *et al.*, who have used rigid cages and fused ring molecules as core structures to which functionalities have been attached combinatorially. T. Carell, E. A. Wintner, A. Bashir-Hashemi and J. Rebek, Jr., *Agnew. Chem. Int. Ed. Engl.*, **33**, 2059-2061, (1994); T. Carrell, E. A. Wintner and J. Rebek, Jr., *Agnew. Chem. Int. Ed. Engl.*, **33**, 2061-2064, (1994); and T. Carrell, E. A. Wintner A. J. Sutherland, J. Rebek, Y. M. Dunayevskiy and P. Vouros, *Chemistry & Biology*, **2**, 171-183, (1995).

Example 5.

Figures 15-18 illustrate an example of library comparison according to a preferred embodiment of this invention. Figure 15 shows a molecule library made up of multiple molecules, and Figure 16 shows a test molecule. Figures 17(a) and 17(b) are a table showing the results of the use of the fragmentation techniques of this invention. The first column of Figure 17 is an arbitrary fragment number. The second column lists fragment size, *i.e.*, the number of atoms in the fragment. The third column is the topological index of the fragment, computed according to the preferred method described above.

Columns four and five relate to the molecule library as a whole. Column four lists the number of molecules in the library in which a particular fragment occurs at least once, and column five lists the total number of occurrences of that fragment in the library. Thus, for example, fragment no. 1 may be found in each of the eleven molecules in the molecule library, giving a total incidence of eleven times in the library--one occurrence for each molecule. The table shown in Figure 17 does not track the number of individual occurrences of each fragment in each molecule in the library or the molecules in which each fragment can be found. This information may be retained in alternative embodiments of the invention.

The sixth column lists the number of occurrences of each fragment in the test molecule.

The last column is a summary of the significance of the fragment comparison process. A fragment is deemed to be significant if it is incident in at least a chosen fraction of the test library molecules. In this case, we have used half of the library molecules as a cut-off, though this number may be the subject of variation in other applications. A "+" is placed in column 7 if a particular fragment occurs in more than half of the library molecules <u>and</u> in the test molecule. A "\" is placed in column 7 if a particular fragment occurs in more than half of the library molecules and does <u>not</u> appear in the test molecule. Nothing is placed in column 7 if the fragment does not appear in more than half of the library molecules. This information can be used for predicting activity of the test molecule based on known activities of the library molecules.

Figure 18 shows the significant fragments identified in this fragment comparison process. Each fragment is labelled by a number denoting corresponding with its entry in the first column of the table in figure 17. Additionally, the fragments are labelled "+" or "\" according to their respective entries in the last column of the table in figure 17.

The kind of information that will come out of any fragment comparison of a test molecule with a molecule library will depend on the characteristics of the molecule library constituents. For example, a molecule library including drugs which show central nervous system activity will provide information useful for predicting the central nervous system activity of a test molecule. As another example, a molecule library including toxins will provide information concerning the toxic activity of a test molecule. Molecule libraries can thus be designed to provide information about multiple activities, both positive and negative, of any test molecule.

<u>Example 6</u>.

In another application of this invention, 500 molecules with some pharmaceutical activity were selected from directories of known drugs. The set contained molecules from different functional classes. Further, 1,500 organic molecules were selected at random from catalogues of three major chemical suppliers. It was ensured that these molecules had molecular weights which fell in the same range as those of the drugs. The ordering of the 1,500 molecules was randomized and divided into three mutually exclusive sets of 500 molecules each. That is, no molecule appeared in more than one set and every molecule from the original 1,500 was present among the three sets. These three sets are taken to represent an "average" sampling of molecules widely used in organic chemistry, covering the multitude of applications that are currently possible.

For each of the four sets of 500 molecules (one set of drugs and three sets of arbitrarily chosen molecules), moieties of size up to 10 atoms were obtained and stored in files along with counters denoting their respective presence (or absence) in each of the molecules in the sets. The moiety and counter files for the drug molecules were compared with those for the three sets of ordinary organic molecules. This exercise enabled detection of moieties which are preferentially present in molecules with pharmaceutical activity as well as a number which are absent. Among the moieties identified were a number already known to medicinal chemists, for empirical reasons. For example, we find that piperidine rings are much more likely to be found in drug molecules than on average in organic chemistry. On the other hand, nitro-groups and nitro-aromatic systems are almost entirely absent from pharmaceutically active molecules, whereas they are found throughout the rest of organic chemistry.

Given that a rigorous, mechanical analysis such as this can confirm long-held intuition of organic and medicinal chemists, it is clear that much more information, which is less immediately obvious, may be derived from this and similar analyses.

Modifications to the invention described above will be apparent to those skilled in the art.

What is claimed is:

1. A process for preparing a database of molecule fragments contained by a molecule, the process comprising the following steps:

(1) selecting a molecule;

(2) identifying substantially all sequentially attached molecule fragments within the molecule;

(3) counting the occurrence of each unique molecule fragment identified in step 2;

(4) storing in a computer-readable storage medium information correlating molecule fragment count with molecule fragment identity.

2. The process of claim 1 further comprising the step of repeating steps 1-4 for a plurality of molecules.

3. The process of claim 2 wherein the plurality of molecules comprise a library of drug molecules, optionally, central nervous system active drugs, or

wherein the plurality of molecules comprise a library of toxic molecules; or

wherein the plurality of molecules comprise a library of molecules chosen randomly from sets of available substances; or

wherein the plurality of molecules comprise a library of molecules exhibiting at least one common activity; or

wherein the plurality of molecules comprise a library of molecules each having a known activity, the activity of the molecules spanning a range of activity values, optionally from low activity to high activity.

4. The process of claim 1 wherein the molecule is described by a two-dimensional representation which permits all atom types and their connectivities to be unambiguously defined, or

wherein the molecule is described by a three-dimensional representation which permits all atom types and their connectivities to be unambiguously defined; and/or

wherein the comparing step comprises the step of arranging the fragments into lists by fragment size.

5. The process of claim 1 wherein the identifying and counting steps comprise the following steps: selecting a fragment and determining whether the selected fragment has been previously stored in a fragment table in a computer-readable storage medium.

6. The process of claim 5 wherein the identifying and counting steps further comprise: repeating the steps of selecting a fragment and determining whether the selected fragment has been previously stored in a fragment table in a computer-readable storage medium until all fragments of the selected molecule below a predetermined threshold fragment size have selected; and/or

wherein the identifying and counting steps further comprise the step of incrementing a counter associated with a fragment stored in the fragment table if the selected fragment has been previously stored in the fragment table; and/or

wherein the identifying and counting steps further comprise the step of storing the selected fragment in the fragment table if the selected fragment has not been previously stored in the fragment table.

7. The process of claim 5 further comprising the step of deriving a first representation of all fragments before the step of selecting a fragment, the determining step comprising comparing the first representation of the selected fragment with first representations of fragments stored in the computer-readable storage medium.

8. The process of claim 7 further comprising the step of deriving a second representation of a selected fragment if the first representation of the selected fragment matches the first representation of a fragment stored in the computer-readable storage medium, the determining step further comprising comparing the second representation of the selected fragment with a second representation of the fragment stored in the computer-readable storage medium; and

wherein the second representation is optionally a graph isomorphism; and/or

wherein the first representation is a topological index; and

wherein the second representation is optionally a graph isomorphism; and/or

wherein the first representation includes atom type; and/or

wherein the first representation includes atom property.

9. A computer-implemented process for predicting activity of a test molecule, the process comprising the following steps:

(1) identifying substantially all molecule fragments within the test molecule;

(2) counting the occurrence of each unique test molecule fragment identified in step 1;

(3) comparing test molecule fragment counts from step 2 with fragment counts for identical fragments from a plurality of molecules having known activity; and

(4) presenting the results of the comparing step as an output.

10. The process of claim 24 wherein the plurality of molecules comprises a library of drug molecules, or

wherein the plurality of molecules comprises a library of toxic molecules, or

wherein the plurality of molecules comprises a library of molecules chosen randomly from sets of available substances, or

wherein the plurality of molecules comprises a library of molecules exhibiting at least one common activity.

11. A process for predicting activity of a test molecule, the process comprising the following steps:

(1) selecting a molecule having a known activity, the molecule comprising molecule fragments;

(2) identifying substantially all molecule fragments within the molecule;

(3) counting molecule fragments identified in step 2;

(4) storing in a digital storage medium information correlating molecule fragment count with molecule fragment identity;

(5) repeating steps 1-4 for a plurality of molecules, each molecule having a known activity; and

(6) selecting a test molecule having an unknown activity, the test molecule comprising test molecule fragments;

(7) identifying substantially all test molecule fragments within the molecule;

(8) counting test molecule fragments identified in step 7;

(9) comparing information correlating test molecule fragment count with test molecule fragment identity with information stored in step 4.

12. A data processing system for creating a database of standardized representations of substantially all fragments of a molecule comprising:

means for selecting a molecule, the molecule comprising molecule fragments;

means for identifying substantially all molecule fragments within the molecule;

means for counting molecule fragments identified by the means for identifying; and

a digital storage medium.

13. A data processing system for identifying a molecule likely to have a desired trait, the system comprising:

means for identifying substantially all molecule fragments within a plurality of molecules;

means for counting in each molecule the molecule fragments identified by the means for identifying;

means for comparing fragment count for one molecule with fragment count for at least one other molecule; and

an output device.

14. A computer-readable medium having recorded thereon a database containing standardized representations of substantially all fragments of a molecule.

15. The computer-readable medium of claim 14 wherein the medium has recorded thereon representations of substantially all fragments of a plurality of molecules.

16. The computer-readable medium of claim 15 wherein the plurality of molecules comprise a library of drug molecules, or

wherein the plurality of molecules comprises a library of toxic molecules, or

wherein the plurality of molecules comprises a library of random molecules, or

wherein the plurality of molecules comprises a library of molecules exhibiting at least one common activity.

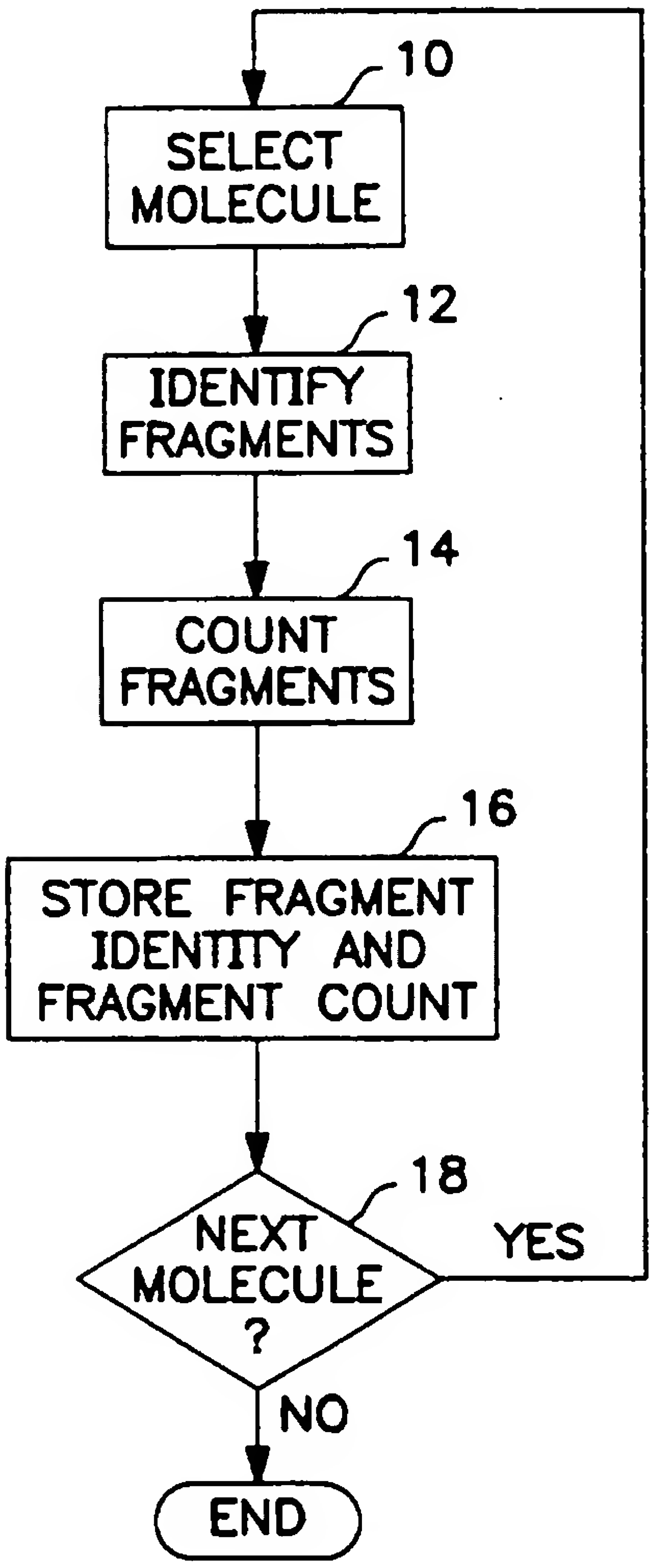

10
SELECT MOLECULE
12
IDENTIFY FRAGMENTS
14
COUNT FRAGMENTS
16
STORE FRAGMENT IDENTITY AND FRAGMENT COUNT
18
NEXT MOLECULE ?
YES
NO
END

FIG. 1

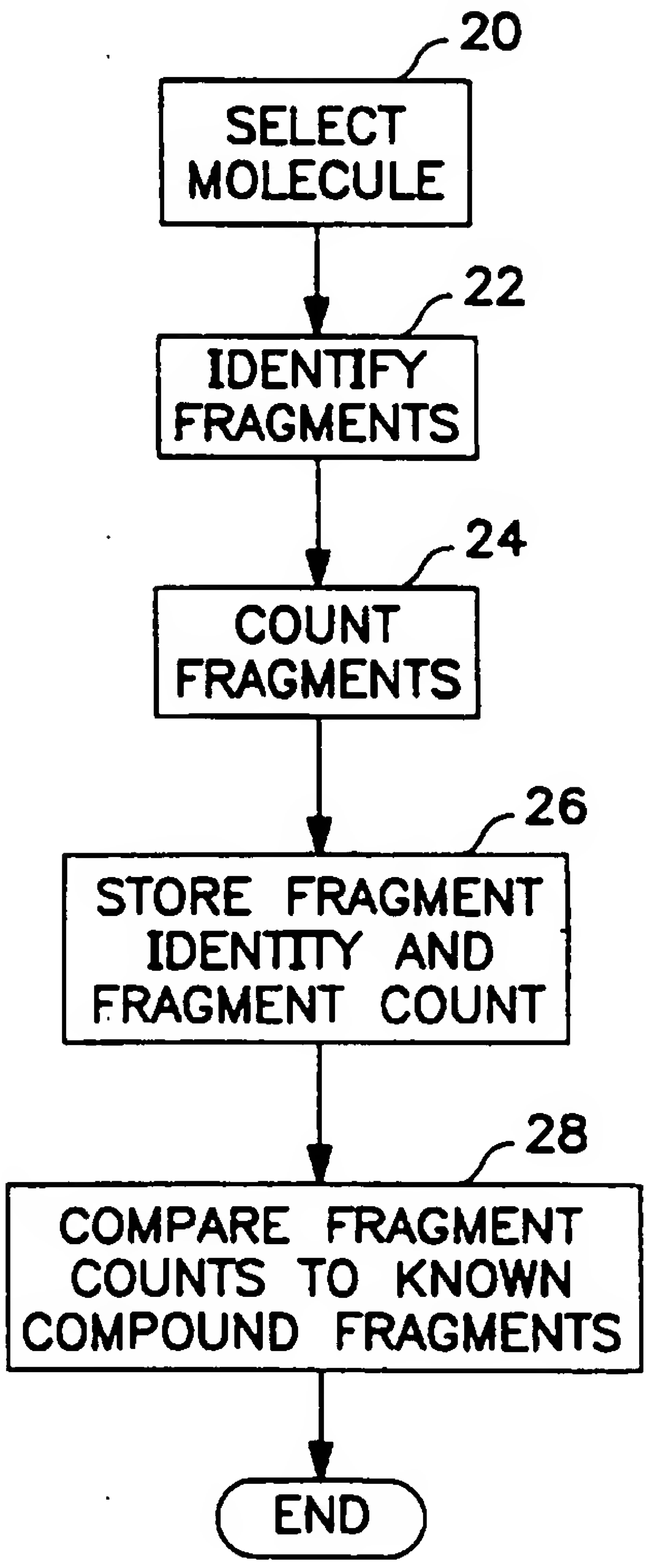
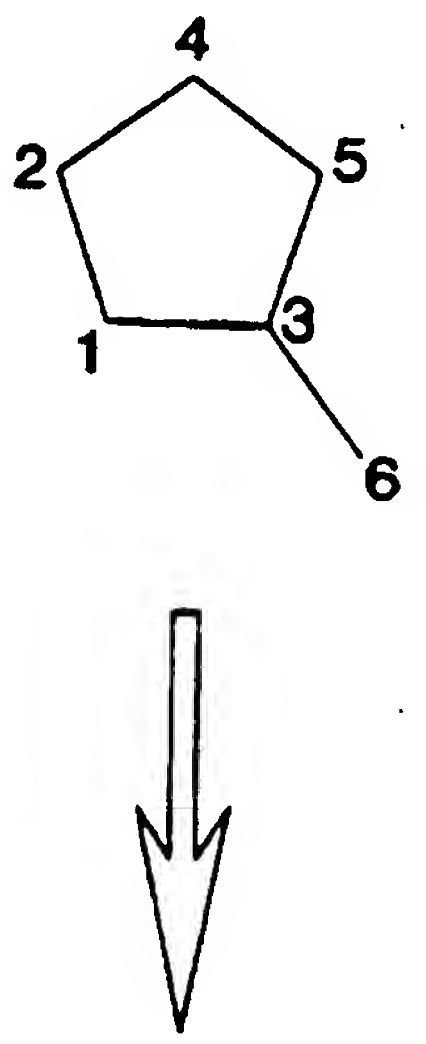

20
SELECT MOLECULE
22
IDENTIFY FRAGMENTS
24
COUNT FRAGMENTS
26
STORE FRAGMENT IDENTITY AND FRAGMENT COUNT
28
COMPARE FRAGMENT COUNTS TO KNOWN COMPOUND FRAGMENTS
END

FIG. 2



Distinct Subgraphs According to Size

1	12	123	1234	12345	123456
2	13	124	1235	12356	
3	24	135	1345	13456	
4	35	136	1356		
5	36	245	2345		
6	45	345	3456		
		356			

Total Number for Each Size

6	6	7	6	3	1

FIG. 3

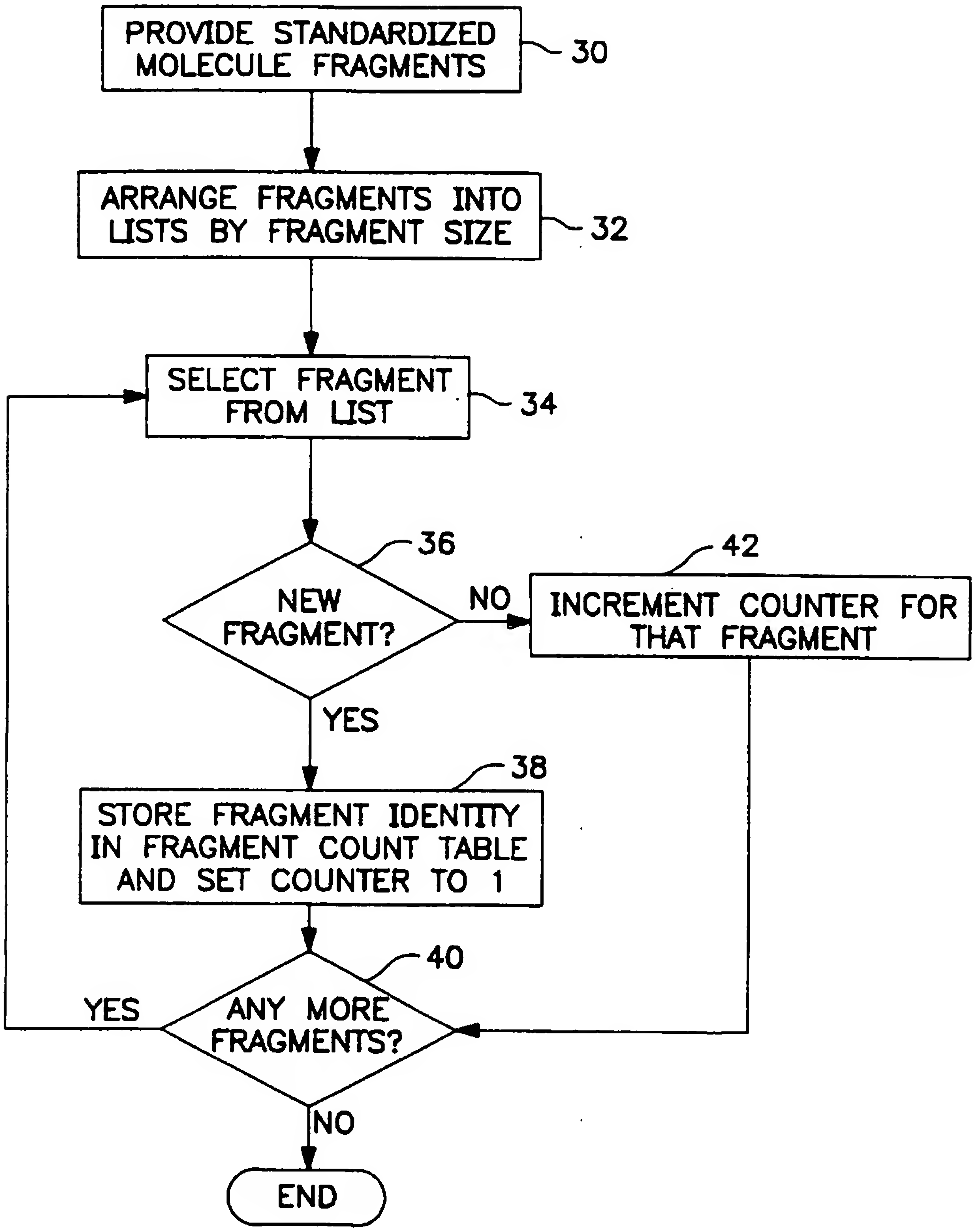

PROVIDE STANDARDIZED MOLECULE FRAGMENTS
30
ARRANGE FRAGMENTS INTO LISTS BY FRAGMENT SIZE
32
SELECT FRAGMENT FROM LIST
34
36
NEW FRAGMENT?
NO
42
INCREMENT COUNTER FOR THAT FRAGMENT
YES
38
STORE FRAGMENT IDENTITY IN FRAGMENT COUNT TABLE AND SET COUNTER TO 1
40
YES
ANY MORE FRAGMENTS?
NO
END

FIG. 4

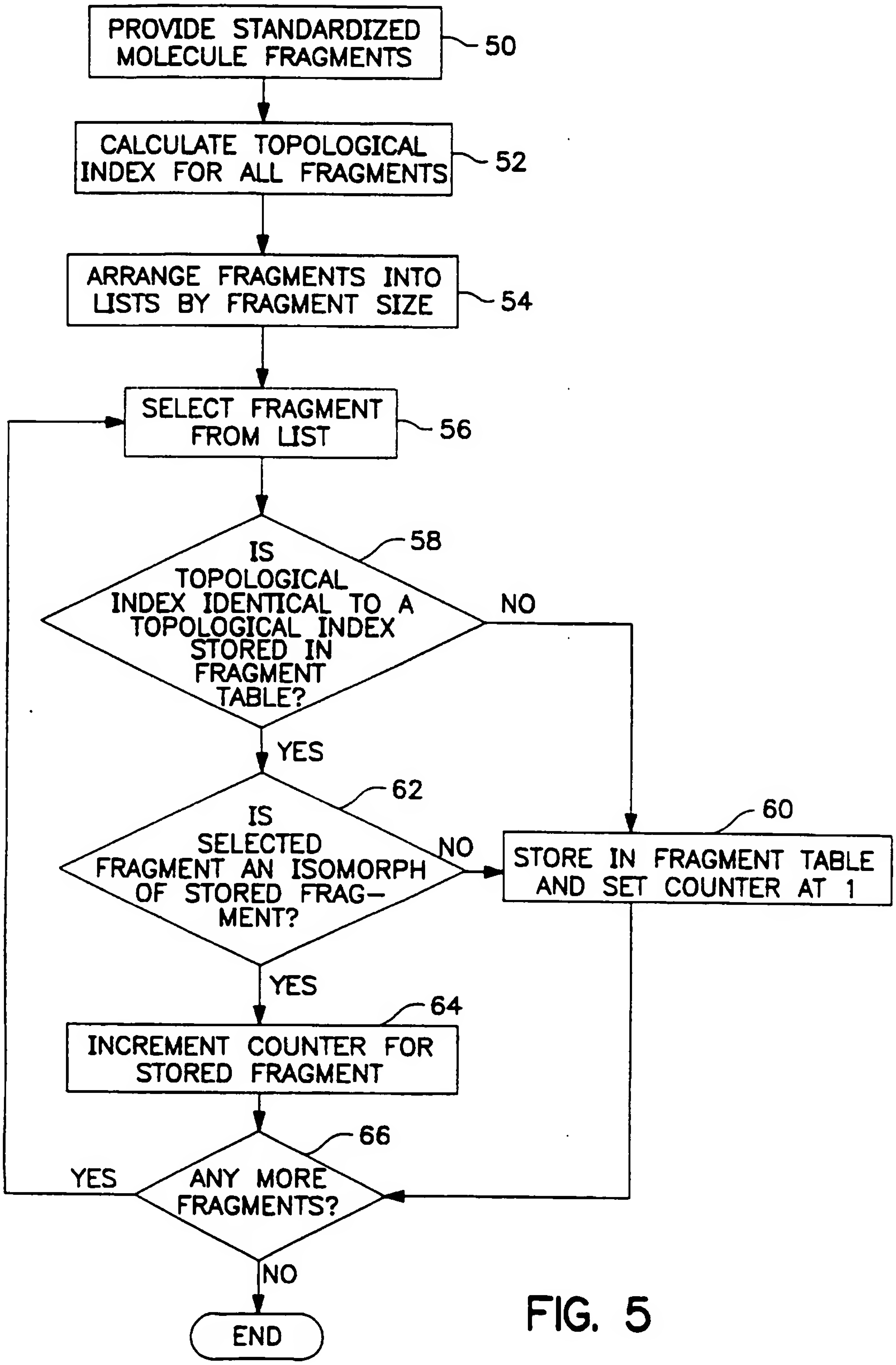

PROVIDE STANDARDIZED MOLECULE FRAGMENTS
50
CALCULATE TOPOLOGICAL INDEX FOR ALL FRAGMENTS
52
ARRANGE FRAGMENTS INTO LISTS BY FRAGMENT SIZE
54
SELECT FRAGMENT FROM LIST
56
IS TOPOLOGICAL INDEX IDENTICAL TO A TOPOLOGICAL INDEX STORED IN FRAGMENT TABLE?
58
NO
YES
IS SELECTED FRAGMENT AN ISOMORPH OF STORED FRAG-MENT?
62
NO
STORE IN FRAGMENT TABLE AND SET COUNTER AT 1
60
YES
INCREMENT COUNTER FOR STORED FRAGMENT
64
ANY MORE FRAGMENTS?
66
YES
NO
END

FIG. 5

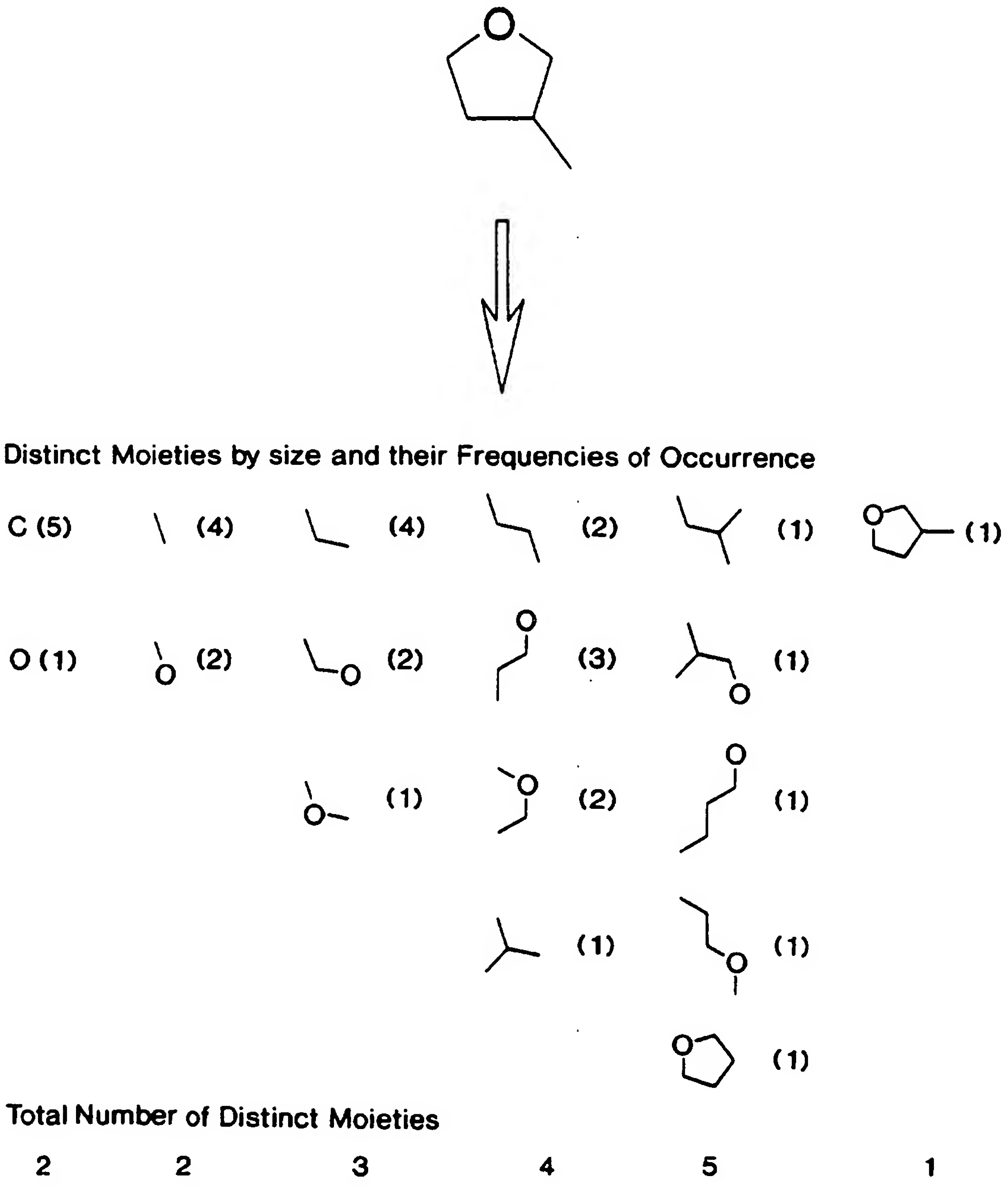

Distinct Moieties by size and their Frequencies of Occurrence
C (5)
(4)
(4)
(2)
(1)
(1)
O (1)
(2)
(2)
(3)
(1)
(1)
(2)
(1)
(1)
(1)
(1)
Total Number of Distinct Moieties
2
2
3
4
5
1

FIG. 6

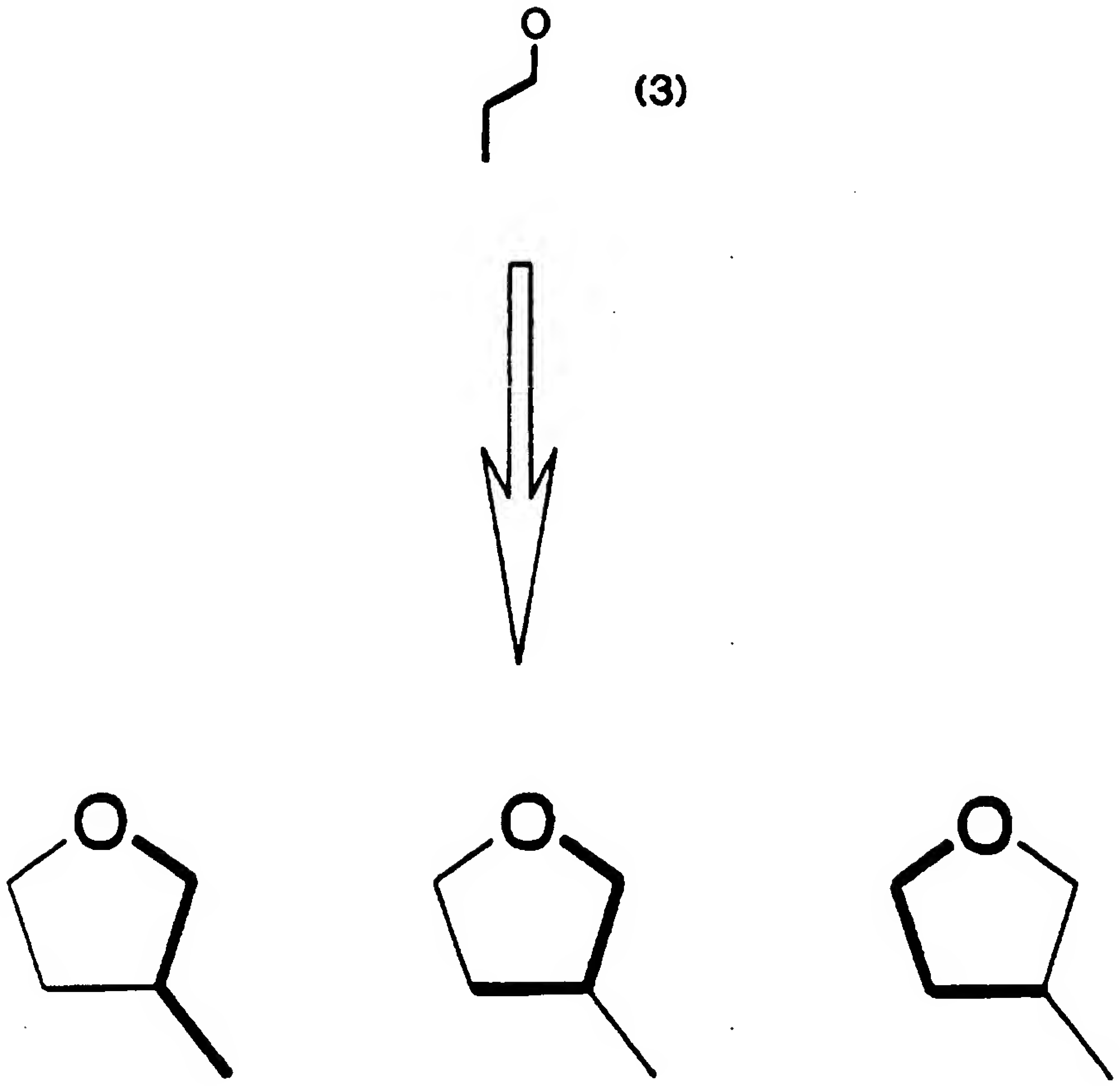


FIG. 7

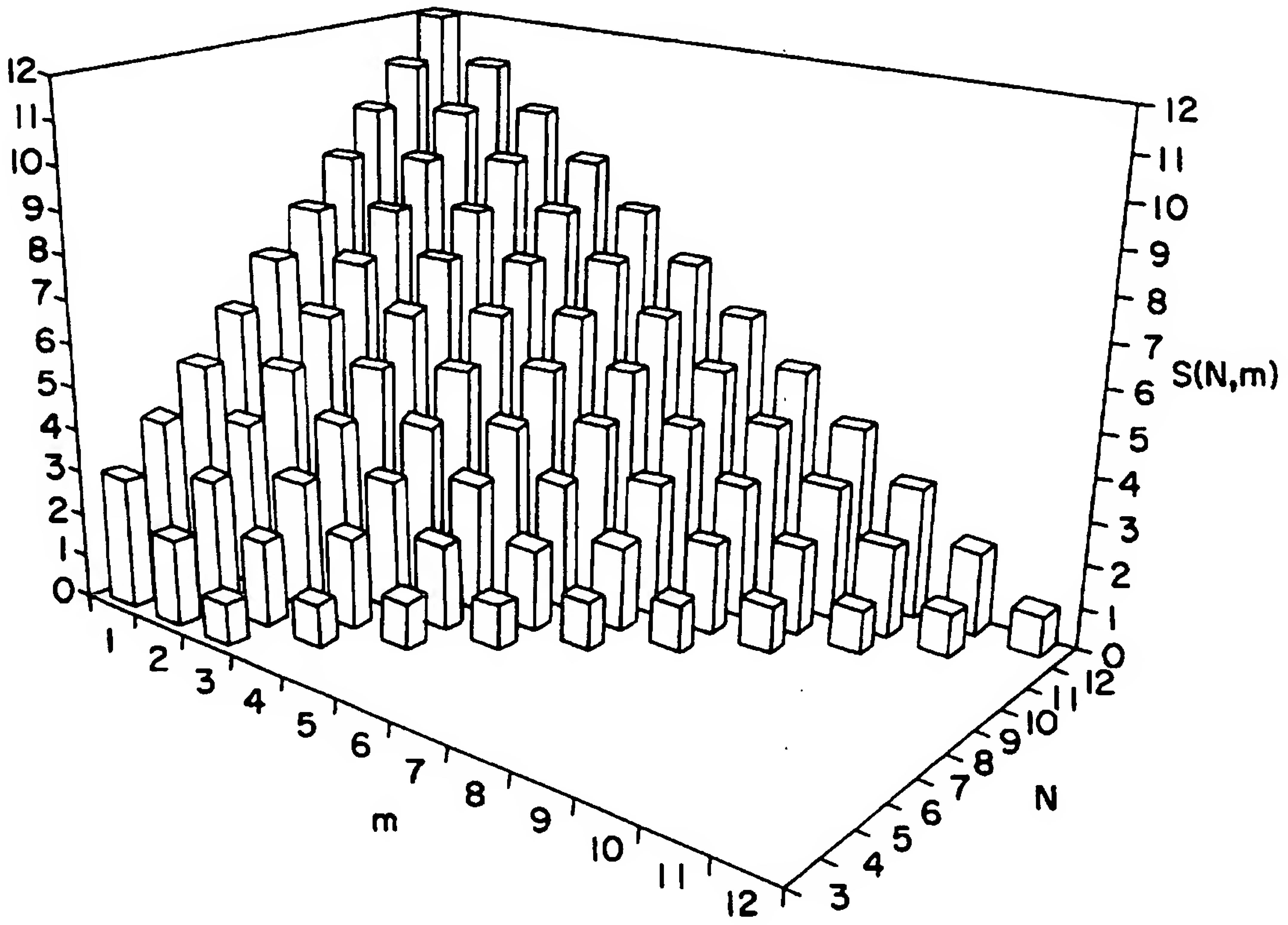


FIG. 8

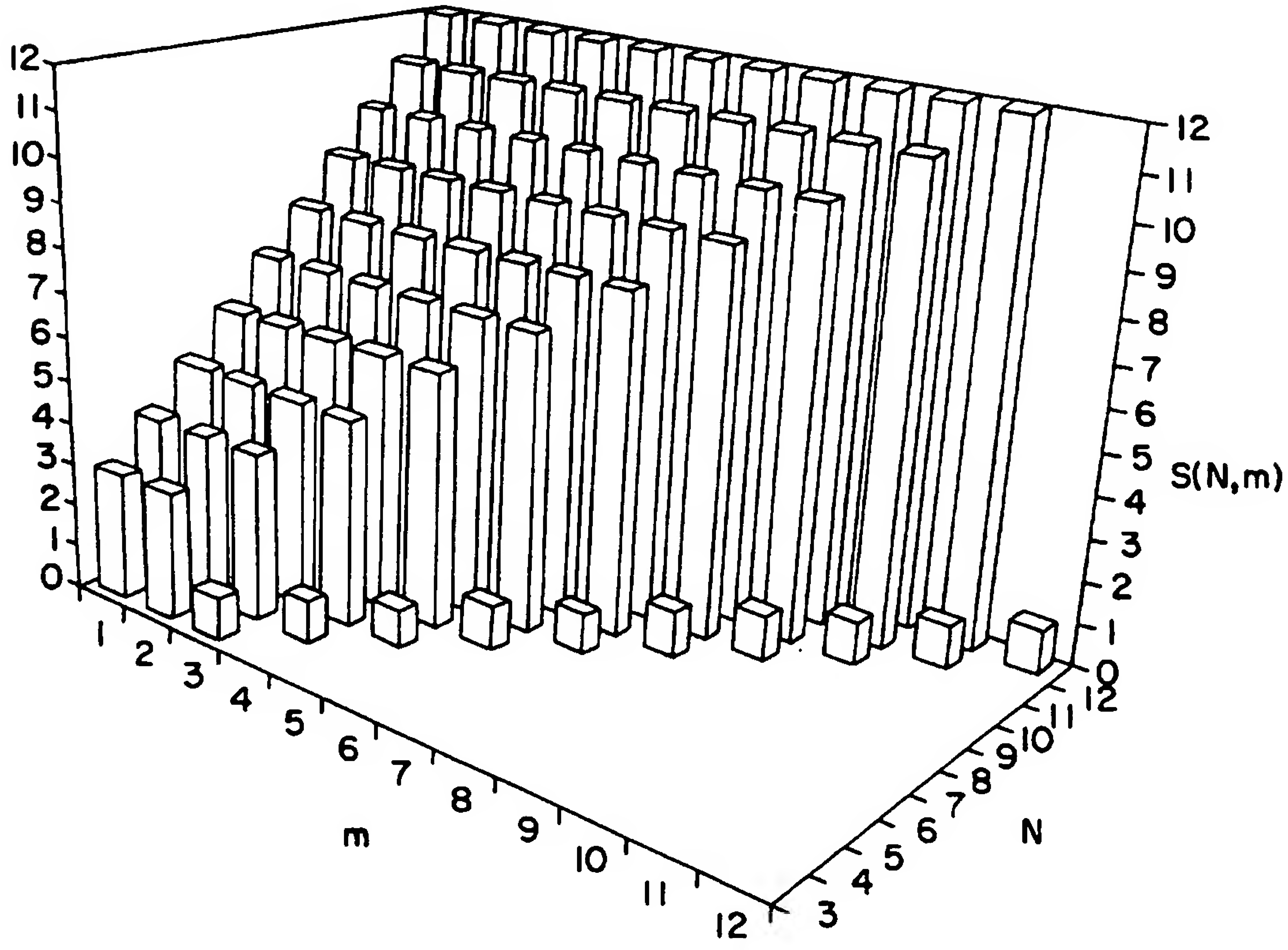


FIG. 9

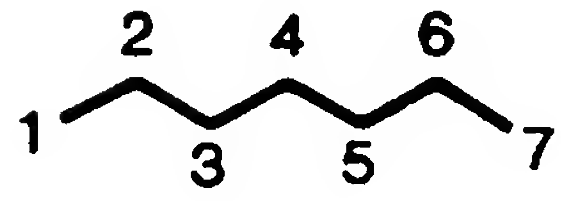


FIG. 10a

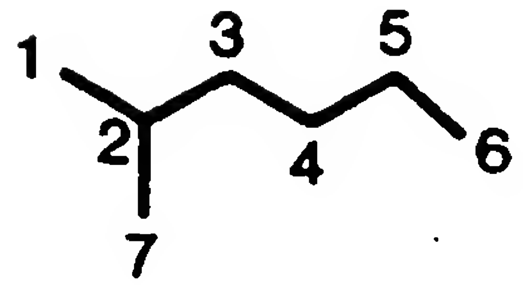


FIG. 10b

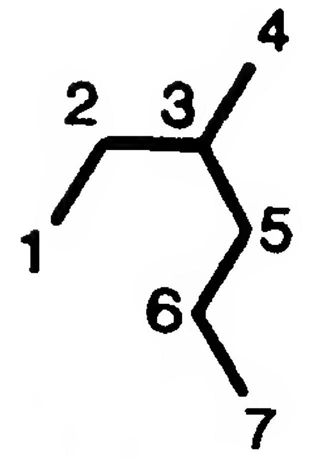


FIG. 10c

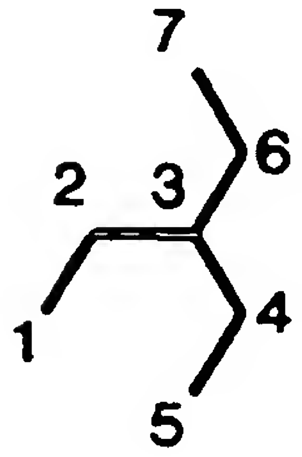


FIG. 10d

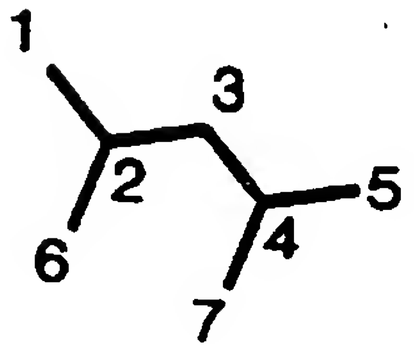


FIG. 10e

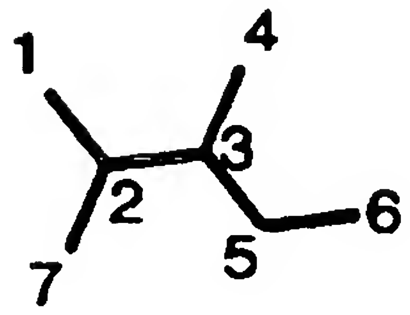


FIG. 10f

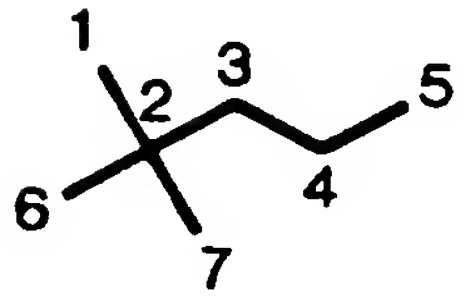


FIG. 10g

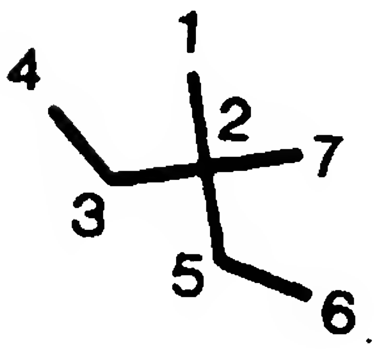


FIG. 10h

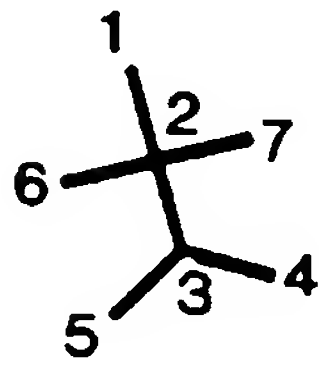


FIG. 10i

		Size (*m*)						
		1	2	3	4	5	6	7
S(7,*m*)								
a	*n*-heptane	7	6	5	4	3	2	1
b	2-methyl hexane	7	6	6	5	4	3	1
c	3-methyl hexane	7	6	6	6	5	3	1
d	3-ethyl pentane	7	6	6	7	6	3	1
e	2,2-dimethyl pentane	7	6	8	8	7	4	1
f	2,3-dimethyl pentane	7	6	7	8	7	4	1
g	2,4-dimethyl pentane	7	6	7	6	6	4	1
h	3,3-dimethyl pentane	7	6	8	10	8	4	1
i	2,2,3-trimethyl butane	7	6	9	11	10	5	1
M(7,*m*)								
a	*n*-heptane	1	1	1	1	1	1	1
b	2-methyl hexane	1	1	1	2	2	2	1
c	3-methyl hexane	1	1	1	2	2	3	1
d	3-ethyl pentane	1	1	1	2	2	1	1
e	2,2-dimethyl pentane	1	1	1	2	3	2	1
f	2,3-dimethyl pentane	1	1	1	2	2	3	1
g	2,4-dimethyl pentane	1	1	1	2	2	1	1
h	3,3-dimethyl pentane	1	1	1	2	3	2	1
i	2,2,3-trimethyl butane	1	1	1	2	2	2	1

FIG. 11

					Size (m)			
Structure/N-position		1	2	3	4	5	6	7
a	1, 7	2	2	2	2	2	2	1
a	2, 6	2	2	3	3	3	2	1
a	3, 5	2	2	3	3	3	2	1
a	4	2	2	3	2	2	1	1
b	1, 7	2	2	2	3	3	3	1
b	2	2	2	3	4	3	2	1
b	3	2	2	3	3	3	2	1
b	4	2	2	3	3	3	2	1
b	5	2	2	3	4	3	2	1
b	6	2	2	2	3	3	2	1
c	1	2	2	2	3	4	3	1
c	2	2	2	3	4	5	3	1
c	3	2	2	3	3	3	3	1
c	4	2	2	2	3	3	3	1
c	5	2	2	3	4	4	3	1
c	6	2	2	3	4	4	3	1
c	7	2	2	2	3	3	3	1
d	1, 5, 7	2	2	2	3	4	2	1
d	2, 4, 6	2	2	3	4	4	2	1
d	3	2	2	2	2	2	1	1
e	1, 5, 6, 7	2	2	2	4	4	3	1
e	2, 4	2	2	3	4	3	2	1
e	3	2	2	3	2	2	1	1
f	1, 7	2	2	2	4	5	4	1
f	2	2	2	3	5	4	3	1
f	3	2	2	3	3	3	3	1
f	4	2	2	2	4	4	3	1
f	5	2	2	3	5	5	3	1
f	6	2	2	2	3	3	3	1
g	1, 6, 7	2	2	2	4	5	3	1
g	2	2	2	3	3	3	2	1
g	3	2	2	3	3	3	2	1
g	4	2	2	3	3	3	2	1
g	5	2	2	2	3	3	2	1
h	1, 7	2	2	2	4	4	3	1
h	2	2	2	2	2	3	2	1
h	3, 5	2	2	3	5	5	3	1
h	4, 6	2	2	2	3	4	3	1
i	1, 6, 7	2	2	2	4	4	3	1
i	2	2	2	3	3	3	2	1
i	3	2	2	3	4	3	2	1
i	4, 5	2	2	2	4	4	3	1

FIG. 12

	m														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
a) *S(N,m)*															
Glycine	5	4	4	3	1										
Alanine	6	5	6	6	4	1									
Cysteine	7	6	7	8	7	4	1								
Serine	7	6	7	8	7	4	1								
Proline	8	8	10	13	13	12	6	1							
Threonine	8	7	9	11	12	10	5	1							
Valine	8	7	9	11	12	10	5	1							
Asparagine	9	8	10	11	12	12	10	5	1						
Aspartic Acid	9	8	10	11	12	12	10	5	1						
Isoleucine	9	8	10	13	15	15	11	5	1						
Leucine	9	8	10	11	12	12	10	5	1						
Methionine	9	8	9	10	10	9	7	4	1						
Glutamic Acid	10	9	11	12	12	12	12	10	5	1					
Glutamine	10	9	11	12	12	12	12	10	5	1					
Lysine	10	9	10	11	11	10	9	7	4	1					
Histidine	11	11	14	18	19	21	25	25	18	7	1				
Arginine	12	11	13	14	14	13	12	12	12	10	5	1			
Phenylalanine	12	12	15	19	24	26	30	34	32	22	8	1			
Tyrosine	13	13	17	22	28	33	37	43	47	42	25	8	1		
Tryptophane	15	16	22	33	48	64	84	102	120	131	122	87	42	11	1

FIG. 13a

	m														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
b) *M(N.m)*															
Glycine	3	4	4	3	1										
Alanine	3	4	5	6	4	1									
Cysteine	4	5	6	8	7	4	1								
Serine	3	4	5	6	6	4	1								
Proline	3	4	6	9	13	12	6	1							
Threonine	3	4	5	9	11	10	5	1							
Valine	3	4	5	9	9	7	4	1							
Asparagine	3	4	6	9	11	12	10	5	1						
Aspartic Acid	3	4	5	8	9	10	8	5	1						
Isoleucine	3	4	5	9	11	12	10	5	1						
Leucine	3	4	5	9	10	9	7	4	1						
Methionine	4	5	7	10	10	9	7	4	1						
Glutamic Acid	3	4	5	8	9	9	10	8	5	1					
Glutamine	3	4	6	9	10	11	12	10	5	1					
Lysine	3	4	5	8	9	9	9	7	4	1					
Histidine	3	6	10	17	19	21	25	25	18	7	1				
Arginine	3	5	8	12	13	13	12	12	12	10	5	1			
Phenylalanine	3	5	6	12	16	24	29	34	32	22	8	1			
Tyrosine	3	5	7	15	20	30	36	43	47	42	25	8	1		
Tryptophane	3	5	8	17	28	52	79	100	120	131	122	87	42	11	1

FIG. 13b

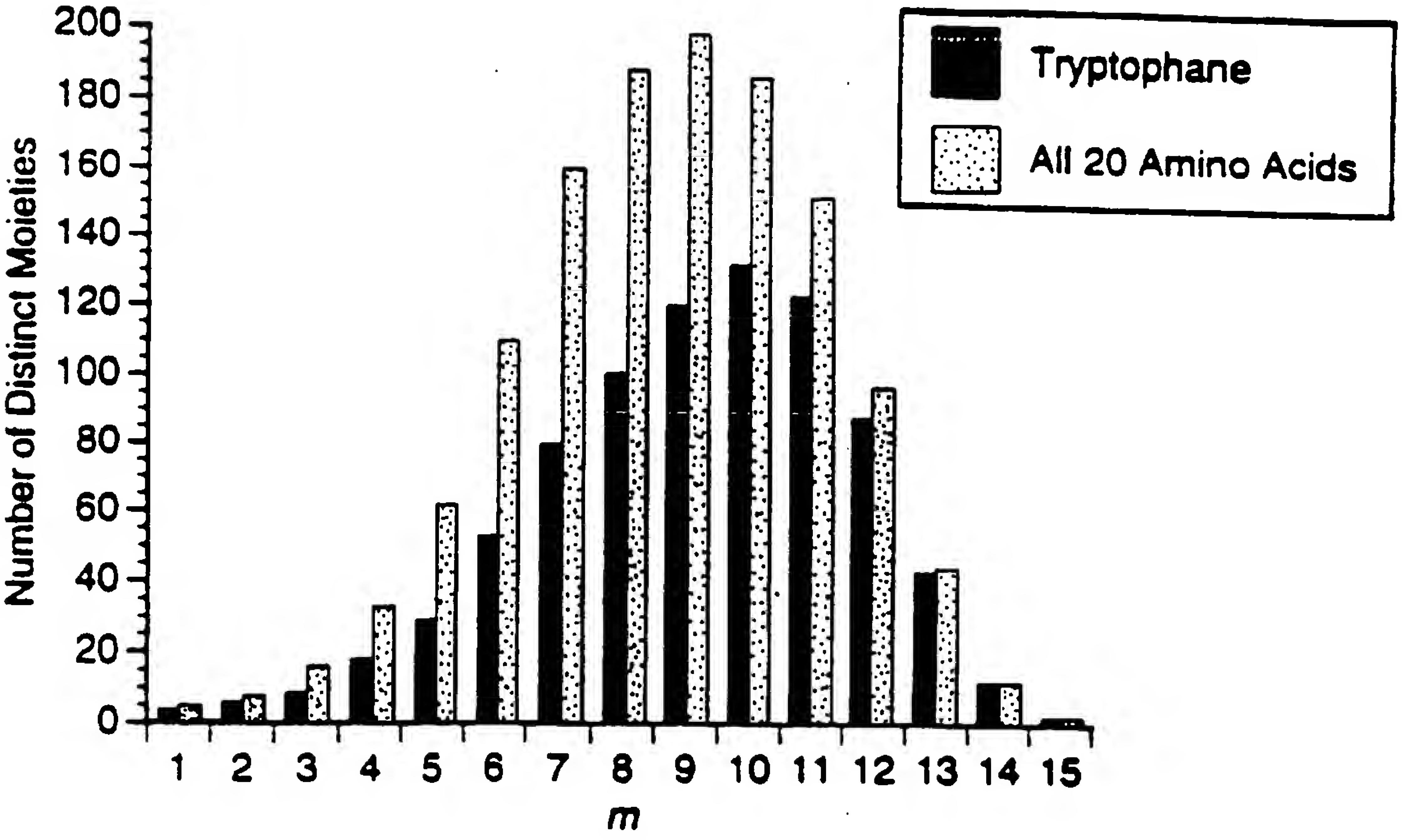

Number of Distinct Moieties
0
20
40
60
80
100
120
140
160
180
200
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
m
■ Tryptophane
▨ All 20 Amino Acids

FIG. 14

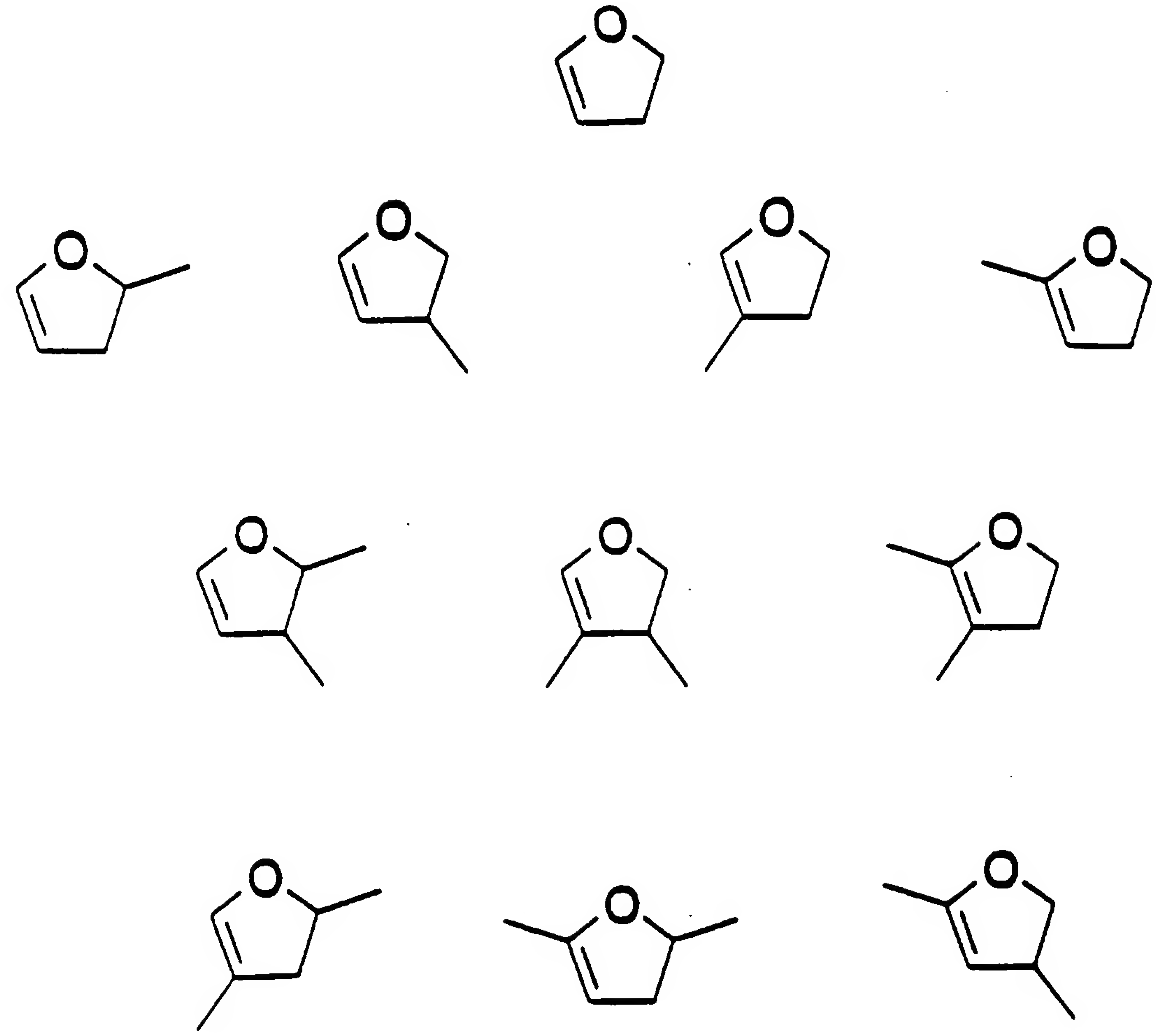


FIG. 15

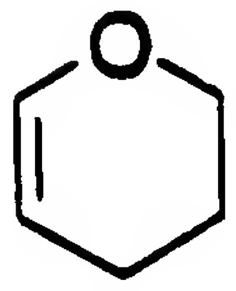


FIG. 16

			Library		Test Molecule	
Fragment #	Size,*m*[a]	Index[b]	Incidence[c]	Count[d]	Count[e]	Selectivity[f]
1	1	0.4082483	11	11	1	+
2	1	0.5000000	11	60	5	+
3	2	0.4472136	11	22	2	+
4	2	0.5000000	11	11	1	+
5	2	1.0000000	11	38	3	+
6	3	0.1825742	11	11	1	+
7	3	0.4082483	11	11	1	+
8	3	0.3162278	11	19	1	+
9	3	0.4082483	11	19	1	+
10	3	0.7071068	11	27	2	+
11	4	0.4824045	11	15	1	+
12	4	0.4023689	11	11	1	+
13	4	0.5773503	11	19	1	+
14	4	0.7236068	11	15	1	+
15	4	0.6969234	11	15	1	+
16	4	1.0937480	4	4	0	
17	4	1.0000000	7	10	1	+
18	4	1.7320508	4	4	0	
19	4	1.2071068	4	4	0	
20	4	0.7352741	4	4	0	
21	4	0.6666667	4	5	0	
22	5	1.0312510	11	11	0	
23	5	0.6220085	4	4	1	
24	5	1.2844571	4	4	0	
25	5	1.2572697	4	5	0	
26	5	1.0773503	4	4	1	
27	5	0.7930217	4	4	1	
28	5	0.9927993	4	4	1	
29	5	1.5764210	4	4	0	
30	5	1.4797192	4	4	0	
31	5	1.2064497	4	4	0	
32	5	0.7051184	4	4	1	
33	5	1.0771602	4	4	1	
34	5	1.3106602	4	5	0	

FIG. 17a

			Library		Test Molecule	
Fragment #	Size,*m*[a]	Index[b]	Incidence[c]	Count[d]	Count[e]	Selectivity[f]
35	5	0.9300565	4	5	0	
36	5	0.9061392	4	4	0	
37	5	0.7814744	4	5	0	
38	5	0.9772839	4	5	0	
39	5	1.3660254	1	1	0	
40	5	1.3535534	1	1	0	
41	5	1.8020951	2	2	0	
42	6	1.5541873	4	4	0	
43	6	1.6391501	4	4	0	
44	6	1.5849946	4	4	0	
45	6	1.4519114	4	4	0	
46	6	1.5854102	1	1	0	
47	6	1.1034497	1	1	0	
48	6	1.4885812	1	1	0	
49	6	1.2537538	1	1	0	
50	6	1.1969234	1	1	0	
51	6	1.7698004	1	1	0	
52	6	1.1350278	1	1	0	
53	6	1.5040966	1	1	0	
54	6	1.3577107	1	1	0	
55	6	0.9247579	1	1	0	
56	6	1.6376965	1	1	0	
57	6	1.7071068	1	1	0	
58	6	1.4545678	1	1	0	
59	6	1.9846204	1	1	0	
60	6	1.6212344	1	1	0	
61	6	1.3290055	1	1	0	
62	6	1.6959942	1	1	0	
63	6	2.0010532	1	1	0	
64	6	1.3848044	0	0	1	

a) Number of atoms in fragment.
b) Topological Index of fragment.
c) Number of library molecules which contain the fragment.
d) Total number of occurrences of the fragment in the library.
e) Total number of occurrences of the fragment in the test molecule.
f) Significance of fragment: if present in at least half of the library molecules and present in ("+") or absent from ("\") the test molecule.

FIG. 17b

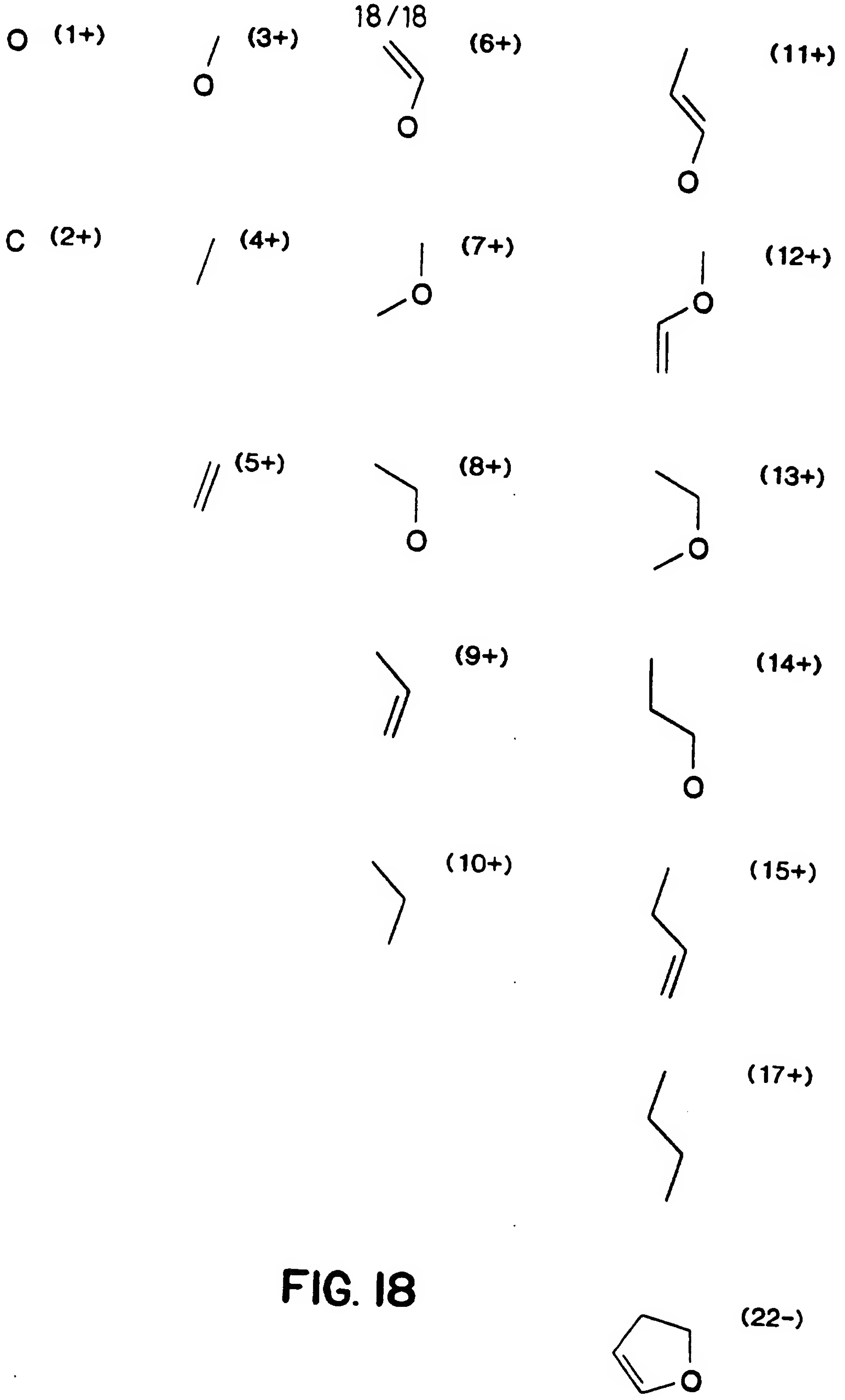


FIG. 18

INTERNATIONAL SEARCH REPORT

Int ional Application No
PCT/US 96/16196

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 G06F17/30 G06F17/50 G01N33/00

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 G06F G01N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X A	EP,A,0 496 902 (IBM CORP) 5 August 1992 see page 3, line 11 - page 4, line 4; figures 1-3	14,15 1-13,16

X	JOURNAL OF CHEMICAL INFORMATION AND COMPUTER SCIENCES, vol. 35, 1995, WASHINGTON US, pages 310-320, XP000576026 SHERIDAN ET AL: "using a genetic algorithm to suggest combinatorial libraries"	14,15
A	see page 312, line 7 - line 19; figure 1	1,9, 11-13

□ Further documents are listed in the continuation of box C.

☒ Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

2

Date of the actual completion of the international search
29 January 1997

Date of mailing of the international search report
06.02.97

Name and mailing address of the ISA
European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl,
Fax: (+31-70) 340-3016

Authorized officer
Guingale, A

INTERNATIONAL SEARCH REPORT

information on patent family members

Int onal Application No
PCT/US 96/16196

Patent document cited in search report	Publication date	Patent family member(s)		Publication date
EP-A-496902	05-08-92	US-A-	5418944	23-05-95
		JP-A-	6028409	04-02-94
		JP-B-	7092804	09-10-95